General Orthopaedics

EOR | VOLUME 4 | MAY 2019
DOI: 10.1302/2058-5241.4.180049
www.efortopenreviews.org

EFORT open reviews

How to prepare and manage a systematic review and meta-analysis of clinical studies

James R. Berstock[1]
Michael R. Whitehouse[2,3]

- Use the PICO framework to formulate a specific clinical question.
- Formulate a search strategy.
- Prospectively register the review protocol.
- Execute the literature search.
- Apply eligibility criteria to exclude irrelevant studies.
- Extract data and appraise each study for risk of bias and external validity.
- Provide a narrative review.
- If appropriate data are available, perform a meta-analysis.
- Report the review findings in the context of the risk of bias assessment, any sensitivity analyses and the analysis of risk of publication bias.
- Useful resources include the Cochrane Handbook, PROSPERO, GRADE and PRISMA.

Keywords: systematic review; meta-analysis; review; orthopaedics

Cite this article: *EFORT Open Rev* 2019;4:213-220.
DOI: 10.1302/2058-5241.4.180049

Introduction

In the 10 years after the 2002 *New England Journal of Medicine* (*NEJM*) publication of a trial that demonstrated no benefit of arthroscopic *versus* sham surgery for knee osteoarthritis,[1] the procedure paradoxically gained popularity and became the most commonly performed orthopaedic surgery in all countries collecting such data. Subsequent trials that confirmed the findings of the *NEJM* publication were conducted, published and largely ignored.[2] It took a review of the totality of the evidence, published in the *BMJ* in 2015,[3] to show beyond doubt that there was no justification for such practice. It is alarming to consider the quantity of medical resources wasted during this period, how they could have been better allocated and the number of inadvertent complications that occurred as a result of this practice.

Evidence synthesis is the bringing together of all the available evidence that meets a predetermined quality assessment and combining, pooling or describing that data with methods appropriate to the type of evidence and data available. Non-systematic reviews are prone to influence by the unconscious or conscious bias of the authors. Systematic reviews offer a thorough appraisal and summary of the evidence and thereby inform clinical practice. They are increasingly seen as providing the most reliable conclusions about treatments. Another important role of systematic reviews is to highlight where no reliable evidence exists (for example only level 3 or 4 studies), thereby directing future research. In this article, we discuss the origins of systematic review and summarize the methodology for performing such work and the meta-analysis which may follow if appropriate data are available.

Pioneering work

The first attempt at a systematic review was performed by James Lind in 1753. In an era when more than half of a ship's crew could perish from scurvy, Lind studied this disease in great detail and was aware of numerous biases existing within the literature. He also realized the importance of a meticulous unbiased review of the 'wealth' of available literature. While serving as the ship's surgeon on board HMS Salisbury in 1747, he trialled different dietary supplements for sailors suffering with scurvy, meticulously controlling all other aspects of the sailors' diets.[4] As a result, he rapidly identified a cure in lemons and, subsequently, limes. He applied this meticulous ideology to his treatise on scurvy:

> '...it is no easy matter to root out prejudices,... It became requisite to exhibit a full and impartial view of what had hitherto been published on the scurvy... Indeed before the subjects could be set in a clear and proper light, it was necessary to remove a great deal of rubbish.'[4]

What is a systematic review?

Chalmers and Altman define a systematic review as a review that has been prepared using a systematic approach to minimizing biases and random errors where the objectives are made clear and the process explicitly documented in the materials and methods section of the review.[5] A systematic review may or may not include a meta-analysis: a statistical analysis of the amalgamated results from independent studies, which generally aims to produce a single estimate of a treatment effect. For a simple clinical meta-analysis, studies must investigate the same intervention and comparator groups and measure the same outcome data or that data be obtainable from the authors.

The key differences between a review article and a systematic review are the explicit steps necessary to demonstrate the elimination of bias. A systematic review must be carried out with the same scientific rigour that is associated with a randomized controlled trial (RCT). There is a need for a very careful consideration of the numerous potential biases that exist within the literature, the direction and magnitude of their effect, and the implications this may have for clinical practice. Ensuring a review is truly unbiased is no small undertaking. For any reader considering undertaking a review, the comprehensive *Cochrane Handbook for Systematic Reviews of Interventions* is available as a free online resource. This article serves as a brief summary.

What is involved in a systematic review?

The systematic review process can be broken down into a few steps. A similar guide by Harris et al exists.[6]

Formulate the question

Initially, this is an iterative process, requiring the reviewer to consider the question to be addressed and the relevant existing literature. The purpose of this is to give the reviewer an understanding of the type of studies that already exist in the literature and enable one to think about what an ideal study might look like in terms of patients, intervention, comparator and outcomes (PICO).[5] Think about the features of methodological quality which could form the basis of inclusion or exclusion criteria. A well formulated concise question with a clear plan from the beginning will save a lot of time and effort later in the process. It is also worth checking to see if a recent systematic review that matches your question has been published or registered.

Develop a protocol

Once a clear review question has been formulated, thought must be given to defining specific inclusion and exclusion criteria for potential studies within a review. Parameters such as the age limits of participants, whether some aetiologies should be excluded and which specific interventions and reported outcomes are of interest should be specified in the protocol. Reviewers also need to consider the level of evidence to be included. Reviews are usually confined to the highest level of evidence available. Ideally multiple RCTs will exist; however, for some clinical questions this level of evidence is unavailable and therefore a review of non-randomized studies may be helpful. It is generally advisable to include all relevant studies, but there may be occasions where clinical practice has changed so radically that historic studies are no longer relevant and should be excluded. It is generally not acceptable to exclude studies based on the language of publication. Having some idea of the quality of evidence available is helpful before writing a detailed and explicit protocol. The protocol should be published before commencement of the review itself. This reduces bias from 'cherry picking' evidence and largely distinguishes a systematic review from a non-systematic literature review. The protocol requires the review team to think carefully about exactly what question they are trying to answer and what type of evidence may bias the overall findings of the review. The PROSPERO International Prospective Register of Systematic Reviews is a quick, free and open-access method of publishing a review protocol (https://www.crd.york.ac.uk/prospero/). Once registered, a systematic review will be assigned a registry number similar to that of a RCT. Review protocols may alternatively be published in journals such as *BMJ Open* and *Systematic Reviews*. Once the protocol has been agreed upon and published, the search for literature may begin.

Literature search

There are several tiers of literature: some is electronically searchable via Medline, Embase, Pubmed, Google Scholar or the trials registries, while some literature exists in journals which are not Medicus Indicus listed. In addition, conference abstracts and proceedings of major journals may not be accessible online (this is known as grey literature). Some studies occur within theses, which are only accessible by visiting university libraries; some evidence may never have been submitted for publication but exists in an even less inaccessible form. The latter category is known as the file drawer problem and is believed to be a larger contributor to publication bias than editors or journals rejecting studies at the review stage. Methods are available for identifying and adjusting for this publication bias which will be discussed later.

Generally speaking, the greater the efforts to identify all the evidence, the more reliable a review is considered. For this reason, multiple databases are often used to search for studies. Embase and Medline would represent the

minimum number of bibliographic databases to search, each containing different subsets of the medical literature. The number of databases required will depend on the nature of the review being conducted and the availability of the evidence. Specific search terms exist for each database; help with the construction of search strategies, including the use of MESH and specific search modifiers, is available online. In addition to hand searching the grey literature, it is generally accepted that searches should not be limited to the English language or just publications from recent years without good reason. Good practice would suggest that personal communications with experts in the field, contacting authors and searching the reference lists of included studies and other relevant articles may unearth additional evidence for inclusion in a systematic review. The searches must be detailed, exhaustive, repeatable and documented in order to find all the relevant published and unpublished studies and to demonstrate reliability. Ultimately, though, a balance must be struck between a truly exhaustive search and the practical constraints of time, cost and other resources. Judgement must therefore be used in setting the appropriate sensitivity and specificity for a particular search strategy. The review team usually collectively agree on the search strategy before eligibility criteria are applied by two independent members of the review team. It is becoming reasonable to assume that the highest quality clinical trials are available in mainstream journals which are all easily identifiable via searches of one or two electronic databases. Identifying lower quality evidence may therefore require more extensive searches.

Apply eligibility criteria

Once the electronic searches have been executed, each potential study must be assessed for eligibility of inclusion against predetermined criteria. This process must be checked by two reviewers acting independently. Formal rules, laid down in the protocol, are required to prevent selective inclusion of studies that support the reviewers' opinions. Disagreements may be settled through arbitration from a third reviewer or, in the case of disagreement when there is not a third reviewer, to include the paper. Many studies can be eliminated by briefly reading the abstract alone, but others require review of the full manuscript or require translation of a foreign language manuscript before eligibility criteria can be applied. Bibliographic managers such as EndNote may be helpful for logging eligibility decisions and managing this part of the process. The numbers of studies at each stage of the search are usually summarized by means of a flow diagram (Fig. 1).

Extract data

To reduce bias, data extraction should also be performed by two reviewers independently, usually recorded onto a standardized proforma for each study included in the review. The type of data to be extracted should be specified in the protocol. This usually includes when and where the study was performed, a summary of the methodology and primary outcome measure used, the number of participants in each intervention group, summary demographic data for each group regarding age, gender, aetiology, outcome measures, complications and length of follow-up.

Assess risk of bias

Bias can be introduced by flawed study designs, inappropriate patient selection and biased treatment group allocation, known and unknown confounders, at the data collection or analysis stages, and, of course, publication bias can skew the available literature.

The Cochrane risk of bias tool provides a standardized guide for considering where bias may creep into a RCT.[6] The categories are very general and include: sequence generation; allocation concealment; masking of participants and assessors; incomplete outcome data; and selective reporting. Randomization is a key part of reducing selection bias and is not always adequate, e.g. quasi-randomization by alternating day of the week on which an intervention is performed. In the case of non-randomized studies, alternative tools such as the Newcastle-Ottawa Scale are available.

The risk of bias assessment described above focuses on internal validity, i.e. the extent to which overall systematic error (bias) is minimized in a clinical study. Consideration now needs to be given to external validity, which is the extent to which the results of a trial offers an accurate basis for applicability to real-life clinical circumstances. For example, a theoretical trial of the lateral *versus* posterior approach for hip replacement may under-report any difference in dislocation rate if patients in both arms of the trial receive constrained acetabular components. In the same example, limiting follow-up to include only the first post-operative days of a strictly enforced period of hip precautions in the inpatient stay may have a similar effect. In both examples, the test was fair but lacked generalizability, known as external validity. The circumstances and delivery of interventions may therefore impact the external validity of an otherwise well-conducted study and need careful assessment.

The assessment of bias is crucial to systematic reviews. Authors must describe where potential bias occurred in an individual study and then make a judgement call to determine whether this would cause a low, high or an unclear risk of bias. Finally, the magnitude and direction each bias may have on the results being reported in a study are also important to consider. For example, inadequate blinding of patients has more of a deleterious effect on subjective outcomes such as satisfaction or functional score than it

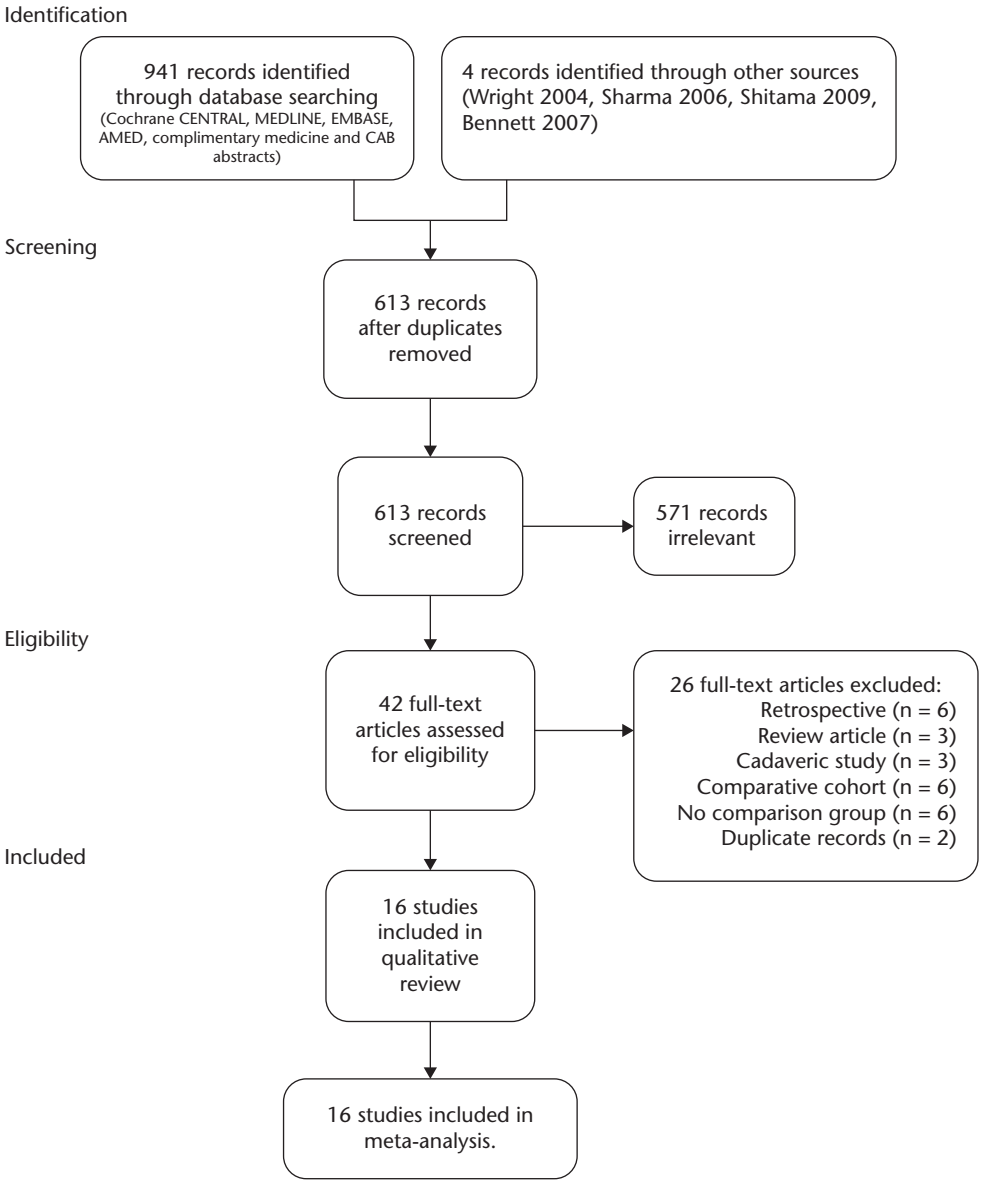

Fig. 1 Example study flow diagram.

does for objective outcomes such as mortality. The risk of bias is usually summarized in a table (Table 1).

Provide a descriptive synthesis

A descriptive synthesis of the studies included in the systematic review with a rigorous attempt to eliminate the effect of biases is challenging but forms the write-up stage of the systematic review. GRADE guidelines are useful for considering the strength of any recommendations or conclusions one may be able to draw.[7] The PRISMA checklist is also useful for demonstrating that the systematic review has followed agreed standards in reporting systematic reviews.[8] It is worth reviewing the PRISMA checklist before commencing the write-up as it provides a logical, comprehensive and reproducible structure to the report.

Meta-analysis

So far, this article has discussed how to take an explicitly unbiased approach to gathering all the information on a topic, critically appraising and synthesizing it to answer a clinical question. When such studies contain data, it is possible to pool results together giving a weighting according to our confidence in them. This is a statistical process known as meta-analysis. Freely downloadable software such as Review Manager can be used to perform meta-analysis and to create forest plots, such as that shown in Fig. 2.

Table 1. Example of a Cochrane risk of bias table

Author (year)	Khan (2012)	Goosen (2011)	Fink (2010)	Shitama (2009)
Random sequence generation	−	−	+	−
Allocation concealment	?	?	+	?
Blinding of participants	−	−	+	?
Blinding of outcome assessment	−	−	−	?
Incomplete outcome data (attrition bias)	−	−	−	−
Selective reporting	−	−	−	−

Grading system: + high risk of bias; ? unclear risk of bias; − low risk of bias

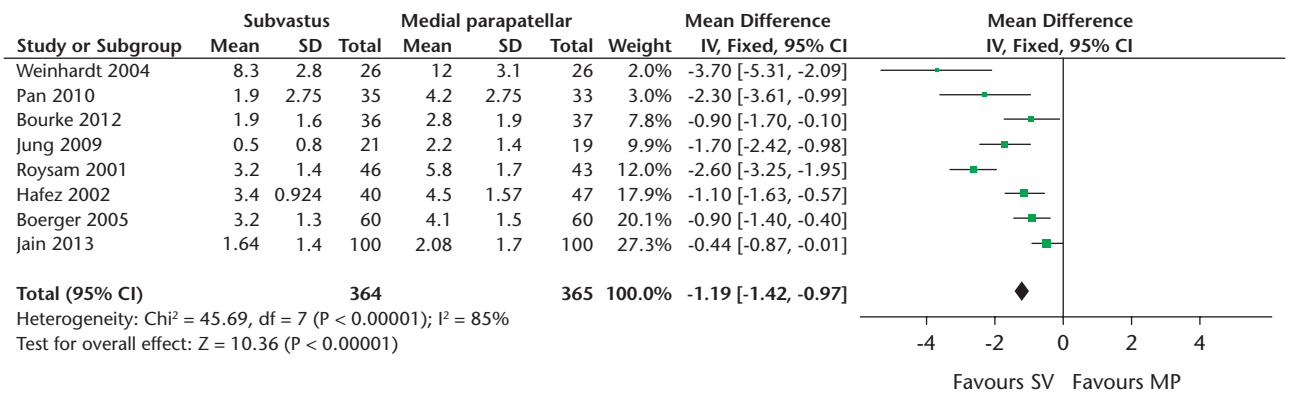

Fig. 2 Forest plot of return of active straight leg raise following either subvastus or medial parapatellar approach to total knee replacement. Pooled data from eight RCTs show that SLR returns 1.19 days earlier with use of the subvastus approach, (95% CI 0.97 to 1.42, p < 0.00001). Data from: Berstock JR, Murray JR, Whitehouse MR, Blom AW, Beswick AD. Medial subvastus versus the medial parapatellar approach for total knee replacement: A systematic review and meta-analysis of randomized controlled trials. *EFORT Open Rev* 2018;3:78–84.

Interpreting a forest plot

A forest plot is a visual representation of the results of a meta-analysis (Fig. 2). Trials are usually ordered by weighting or year of publication. Each square corresponds to the risk or odds ratio, the size of the square to the sample size and the horizontal lines correspond to the 95% confidence intervals (CI). The centre of the diamond represents the pooled relative risk or odds ratio with its 95% CI denoted by its width.

The headline results of a meta-analysis are usually denoted by a summary mean difference or risk ratio, a 95% CI and a test of overall statistical significance expressed as a p-value. These three values indicate the magnitude of the treatment effect, the spread and the statistical significance of the effect. These data are most useful but should not be considered individually. For example, pooling of large amounts of data may produce very small p-values, suggesting a strongly statistically significant result; however, the actual size of the effect may be small and clinically of little relevance. Confidence intervals may also indicate that the true treatment effect may be harmful, despite the mean pooled effect indicating benefit. This would usually indicate that more studies are required in order for a definitive conclusion to be drawn.

Statistical assumptions

Broadly speaking, there are two different assumptions on which meta-analysis can be performed. These are the fixed and random-effects models. A fixed effects meta-analysis assumes that all studies are measuring the same underlying treatment effect and, if it was not for random (sampling) error, all results would be identical. In this model, larger studies tend to get a larger weighting. This is usually an advantageous assumption when there is low heterogeneity between studies.

The random-effects model assumes that the true treatment effect actually varies between studies to form a normal distribution of effect sizes. When significant inter-study heterogeneity exists, this is usually the preferable model.

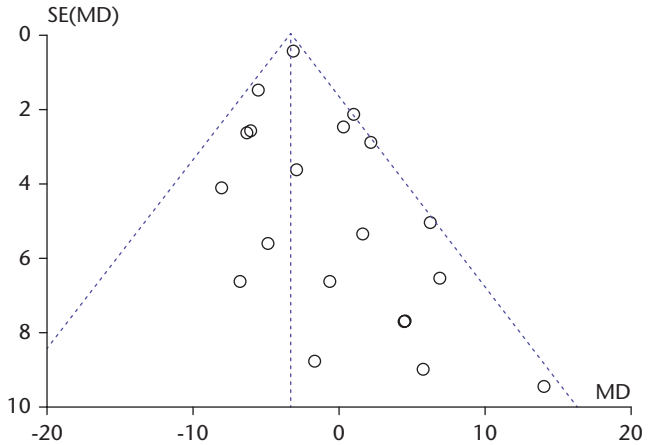

Fig. 3 A theoretical funnel plot demonstrating an absence of small studies reporting a negative effect, suggestive of a publication bias. The central vertical line shows the pooled effect size and each circle represents a study. Larger studies with a smaller standard error are located at the apex of the triangle. Smaller studies are expected to distribute on both sides of the overall effect size line, but in this example do not.

However, it has the effect of increasing the weighting of smaller studies and accordingly decreases the weighting of larger studies. This may not be desirable if the smaller studies are thought to be methodologically less reliable.

Publication bias

Funnel plots can be used to help detect publication bias (Fig. 3). Publication bias arises from the way researchers and journal editors have a tendency to handle positive and negative study findings differently. A funnel plot is essentially a graph of the inverse of the standard error of a study on the y-axis and individual study effect size on the x-axis. If enough studies exist, a triangle or inverted funnel will emerge if there is no evidence of publication bias. The large studies with small standard errors accumulate near the top of the chart, close to the pooled effect size, while the small studies are usually subject to larger sampling error and display a wider spread of results around the pooled effect size. When part of the triangle is missing, there is said to be asymmetry, which may indicate publication bias.

Usually, in the presence of publication bias, studies in the low left-hand part of the funnel are absent. These are small, negative studies. The 'trim and fill' method may be used when there is asymmetry in the funnel plot. This corrects the results of a meta-analysis in the face of a publication bias. This method also estimates how many studies are missing as a direct result of publication bias.

Sensitivity analysis

Following meta-analysis, various assumptions can be tested to identify the robustness of the pooled effect size. These are known as sensitivity analyses and should be

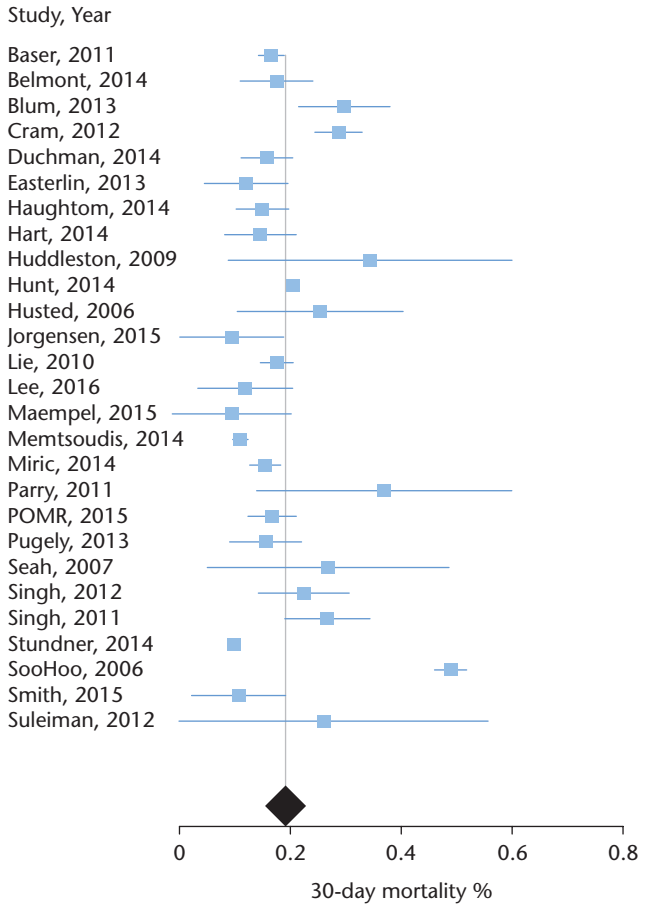

Fig. 4 Simplified forest plot of 30-day mortality following total knee replacement. Pooled data from 1.8 million total knee replacements, 30-day mortality 0.19% (95% CI 0.15 to 0.23).

defined *a priori* in the study protocol rather than on a *post hoc* basis. Usually heterogeneity is examined. Heterogeneity is the difference between individual study findings that is not due to chance. Clinical heterogeneity is often present, e.g. small differences in the delivery of the intervention and differing patient demographics or response to treatment. Clinical judgement must be used to postulate the cause of heterogeneity.

Quantifying heterogeneity may be useful. The I^2 test represents the difference between observations in studies that would not be expected by statistical chance alone. Conventionally, $I^2 < 25\%$ signifies low heterogeneity, whereas $I^2 > 75\%$ indicates a high level of heterogeneity.

The potential causes of such heterogeneity need careful consideration. Heterogeneity may be due to patient diversity, different treatment effects or study bias, and can be explored using subgroup analysis, meta-regression or a funnel plot. For example, cultural, temporal and geographical differences in social care and expectations of length of stay following arthroplasty surgery may contribute to a wide variation in worldwide studies reporting this

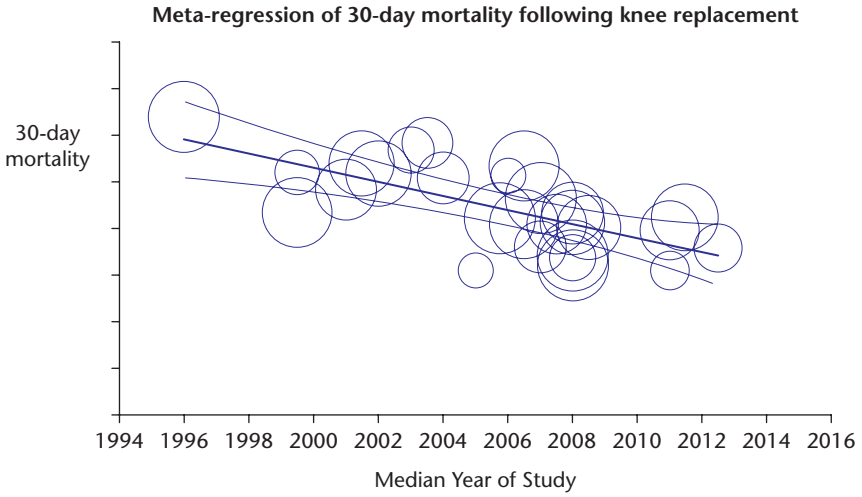

Fig. 5 Meta-regression of data presented in Fig. 4. Each individual study is represented by a circle, the size of which correlates to the weighting of each study. The straight line shows that mortality is seen to decrease with time, with the lines above and below representing 95% CIs, which get tighter where more data exist. For this meta-regression R2 = 70%, suggesting that 70% of the variation between studies is explained by the year of data collection. Data with permission from: Berstock JR, Beswick AD, López-López JA, Whitehouse MR, Blom AW. Mortality after total knee arthroplasty: a systematic review of incidence, temporal trends, and risk factors. *J Bone Joint Surg Am* 2018;100:1064–1070.

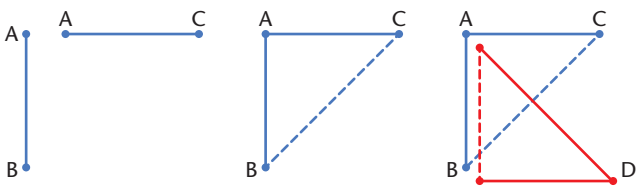

Fig. 6 Schematic of comparisons possible via network meta-analysis.

as an outcome measure. Subgroup analyses of studies grouped by geographic region or decade of study might produce more homogenous results, therefore explaining the observed heterogeneity supporting the initial hypothesis. While subgroup analyses are good for categorical covariates, meta-regression is often used to explore relationships between continuous covariates. Meta-regression could for example be used to explore a temporal trend towards shorter lengths of stay that could not be investigated in any of the original studies.

Meta-regression

Meta-regression is the meta-analysists equivalent of simple or multiple regression in cohort studies. Additional co-variates may be studied for their effect within a meta-analysis. For example, studies reporting 30-day mortality following total knee replacement can be displayed in a forest plot (Fig. 4). Meta-regression can be used to determine the effect of a co-variate, such as median year of data collection on the overall meta-analysis (Fig. 5). Study weighting is retained in meta-regression, whereas in simple regression there is no weighting.

Individual patient data meta-analysis

Individual patient (or participant) meta-analysis is a subtype of the meta-analysis technique. Rather than being reliant on the summary aggregate data typically reported in publications describing RCTs, standardized patient level trial data are sought from authors to create a combined common dataset across multiple trials. This allows for re-analysis of the data to provide an overall estimate of the treatment effect of an intervention. This method is very resource-intensive and difficult to do but can offer novel and definitive answers that may be missed in the analysis of summary data alone. For this reason, the Cochrane Methods Group recognizes this as the 'gold standard' of systematic review.[9]

Network meta-analysis

Network meta-analysis is a method for establishing the treatment effect of interventions not directly compared when there are appropriate randomized studies available with a common comparator. This method is typically more representative of clinical decision-making regarding the treatment of patients where there may be multiple treatment options. The common comparators are defined and networks formed to establish where effects can be assessed between different interventions (Fig. 6). Where there are two RCTs, the first comparing intervention A and intervention B and the second comparing intervention A and intervention C with common outcomes reported, combining the data from these studies allows clinically relevant estimates to be made of the differences between interventions B and C. These networks are then built up in increasing layers of complexity.

Conclusions

In a world where medical knowledge is being produced at a bewildering rate, the aggregation and rigorous appraisal of the totality of available evidence will become increasingly important if such information is to be made usefully available to clinicians. Evidence synthesis needs to be performed in order to gather knowledge, inform policy, process and practice, establish the need or the lack of need for research in an area and to generate hypotheses. The role of systematic review will therefore be increasingly important to the future conduct of evidence-based medicine. It is also increasingly unlikely that researchers will achieve funding of a study without an adequate and robust systematic review of the topic to inform the need for research in the area and some journals require a brief systematic review to accompany all submissions. This article serves as a summary of the steps required to perform a systematic review. We would encourage potential reviewers to refer to the online *Cochrane Handbook for Systematic Reviews of Interventions* for a more detailed description.

AUTHOR INFORMATION

[1]University of British Colombia Department of Orthopaedics, Gordon & Leslie Diamond Health Care Centre, Vancouver, British Columbia, Canada.

[2]Musculoskeletal Research Unit, School of Clinical Sciences, University of Bristol, Southmead Hospital, Bristol, UK.

[3]National Institute for Health Research Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust and University of Bristol, Bristol, UK.

Correspondence should be sent to: James Berstock, University of British Colombia Department of Orthopaedics, Gordon & Leslie Diamond Health Care Centre, 3rd Floor, 2775 Laurel Street, Vancouver, British Columbia, V5Z 1M9, Canada.
Email: jberstock@gmail.com

FUNDING STATEMENT

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

ICMJE CONFLICT OF INTEREST STATEMENT

JRB reports that he works for North Bristol NHS trust as an Orthopaedic Surgeon and is not employed elsewhere.

MRW reports that he has received a grant from National Institute for Health Research (NIHR) and that this study was supported by the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. He reports he has received a grant from Stryker investigate the outcome of the Triathlon total knee replacement. He has received payment for lectures including service on speakers' bureaus from Heraeus, DePuy and that his institution receives payment at market rates for teaching on basic science and cemented joint replacement delivered to trainees and consultants.

REFERENCES

1. **Moseley JB, O'Malley K, Petersen NJ, et al.** A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N Engl J Med* 2002;347(2):81–88.

2. **Siemieniuk RAC, Harris IA, Agoritsas T, et al.** Arthroscopic surgery for degenerative knee arthritis and meniscal tears: a clinical practice guideline. *BMJ* 2017;357:j1982.

3. **Thorlund JB, Juhl CB, Roos EM, Lohmander LS.** Arthroscopic surgery for degenerative knee: systematic review and meta-analysis of benefits and harms. *BMJ* 2015;350:h2747.

4. **Lind J.** A Treatise of the Scurvy. In three parts. Containing an inquiry into the nature, causes, and cure, of that disease, etc: Edinburgh; 1753.

5. **Chalmers I, Altman DG.** *Systematic reviews*. London: BMJ Publishing Group, 1995.

6. **Harris JD, Quatman CE, Manring MM, Siston RA, Flanigan DC.** How to write a systematic review. *Am J Sports Med* 2014;42(11):2761–2768.